

The Invisible Web: What Search Engines Can't Find and Why

Presentation by Laura Gordon-Murnane, MLS
lgordonm@bna.com

University of Maryland Libraries Digital Dateline Series
November 5, 2003

Expectations for Searching the Web

According to the report released by the Pew Internet & American Life Project "[Counting on the Internet](#)" the Internet has become a "mainstream information tool."

Americans rely on Internet to find information on:

- Health Care information,
- Government information,
- News
- E-Commerce

"Search Engines have become the primary starting point for Internet navigation and electronic commerce..."
(WSJ, Nov 3, 2003)

How do people find what they want on the Net?

Pew Internet Report

Findings:

- Search Engines - are indispensable for Internet Users. More than 8 in 10 American Internet users use search engines to find information on the net.
- Search Engines are the most popular way to locate a variety of information.

Are they Successful in Finding the Information they Need?

iProspect [Study](#) examined search engine behavior

Findings:

- 3/4 of Internet users use search engines,
- 16 percent of Internet users look at the first few search results, 32 percent will read through to bottom of the first page,
- 52.1 percent Internet users choose the same search engine or directory when searching for information
- 35 percent will choose an alternative search engine
- 7.5 percent of Internet users refined their searches with additional keywords when results were not what they wanted,
- Less than half (45.9 percent) felt that their searches were successful almost all of the time,
- 1/3 of the survey participants indicated a success rate of three-quarters of the time, and 13.3 percent found what they were looking for half of the time.

Conclusions:

- Expectation is that the information is on the Internet
- Search engines will help you find the information you want.
- Success rate of these searches is less than 50 percent.
- 13.3 percent of those in the survey found what they were looking for half of the time.

Definitions of the Visible and Invisible Web

Definition 1: Visible Web

The Visible Web consists of material found “on the web” that general search engines (Google, Altavista, AlltheWeb, Teoma, etc.) can find and make searchable and easily accessible.

Definition 2: Invisible Web

The Invisible Web consists of material found “on the web” that general search engines (Google, Altavista, AlltheWeb, Teoma, etc.) cannot, will not, or do not crawl/index/make searchable and easily accessible.

“On the Web” vs. “Via the Web”

Information on the Web - Features

- Anyone with server access can place just about anything they want on the web
- Little bibliographic control
- No language control
- Quality of information
- Cost is free or low

Examples of [Yahoo](#), [Firstgov.gov](#)

Information “Via the Web” Features

Using an Internet connection and browser to access traditional/commercial databases and resources

- Databases not directly searchable via web search tools
- Information is highly structured and well indexed
- Quality of information is uniformly high - professional resources
- Invisible web materials can be free, low cost, or expensive
- Proprietary-Cost can vary, often expensive
- Types of materials can include full text of peer reviewed journals, Indexed, abstracts

Examples: [Dialog](#), [Lexis-Nexis](#), [Factiva](#)

How do “on the web” Search Engines Work?

Search Engines Consist of Three Parts

- The Web Crawler
- The Indexer
- The Query Processor

Myth - All Search Engines are Alike

- All search engines are not the same
- Comprehensiveness, currency, and coverage
- Interface, syntax, capabilities - all unique
- Different Algorithms - leads to different results
- If you don't find what you need - consider using invisible web tool
- Learn more than one search engine

The Web Crawler identifies pages in 2 ways:

- Add url form
- Harvest hypertext links embedded on the page

Points to consider

- Harvesting of links creates a very large pool of pages to visit. The crawler must determine if the page has already been visited and if it is already in the search engines's index. If the url is already in the index, the crawler has to determine if the information is still current or if the information in the search engine's index is out of date and needs to be updated.
- Crawling the web is resource-intensive
- Do not assume that every search engine will crawl and index the site's entire set of pages

Search Engine Indexes

- Indexes every word on every page and stores in a huge database.
- The search engine stores the full text of the pages - and this allows for the search engine to offer more than just simple single keyword matching.
- Offer proximity searching that will match multi-word phrase, sentence, and bigger sections of text.

The Query Processor

- Most complex piece of the search engine
- Query Processor has three parts - search form, the engine that processes the request, the results page.
- Search form and results page are similar for all web search engines.
- Key difference between search engines - the way relevance is calculated.
 - Statistical Analysis of Text
 - Link Analysis
 - Clustering

Myth - Search Engine Indexes are Current

Search Engines search their index - not the current web

Crawling the web is resource intensive - the search engine has to determine how frequently it will recrawl the page.

Many search engines have increased their recrawl rates

Web is too large for any one search engine to provide comprehensive coverage. Too many new pages, too expensive to recrawl every new page. This is for the Visible Web.

What about the Invisible Web. Why can't crawlers find Invisible Web Pages?

Why Search Engines Can't Find Pages

Technical Issues - Invisible Web

Opaque Web

- Disconnected Pages - no links on the page
- Page not submitted to an engine/engines (this is a secondary way engines learn of new content)
- Depth of crawl - the crawler does not crawl the entire site. (tip: use the search engine at the source for better search results)
- Maximum number of viewable results
- Size limitation - each engine decides how much of the page it will crawl.
 - Google only indexes the first 101k of the page
 - AllTheWeb - indexes the entire page
 - Altavista - indexes the first 110 k of the entire page
- Frequency of Crawl - Content on the page is not current
 - Each engine is different
 - Many pages have a minimum of 30-45 days after discovery
- Recrawl rates - Each engine is different

Private Web

- [Robots Exclusion Protocol](#) (Don't crawl and index my content)
 - Noindex Meta Tag (specifies specific page or pages the crawler is not supposed to crawl).
 - Firewall is in place - only those authorized to gain access are allowed in
 - Password protected pages
-

Proprietary Web

- Password protection
 - Firewall - Only those authorized to gain access are allowed in
 - Use legacy database systems - available long before Web existed
-

Invisible Web

- Non-html text - Search engines are designed to index html text - audio, video, images - these formats are hard for search engines to understand - many search engines now include searching for non-text files
 - Multiple Formats - Not every format is crawled by every search engine - pdf, flash, shockwave, executable programs, compressed files - technically indexable - until recently ignored by search engines. To index these formats is resource-intensive. ([Google](#), [Altavista](#), [AlltheWeb](#), [Firstgov.gov](#) - are now including non-html format options). Take a look at [CiteSeer](#) (Scientific Literature Digital Library - NEC Research Institute) - indexes pdf files - also creates a citation index - easy to locate related documents).
 - Codes and Frames are difficult for [Web Search Engines](#)
 - Registration Forms - the form cannot be completed by the spider - blocking access to the information.
 - Relational Databases
 - Why do web content developers use databases?
 - Flexibility, easily maintained
 - Web front ends to provide access to proprietary systems that are now open - these databases are available "via the web"
 - Dynamically Generated Pages - search engines [refuse](#) to crawl any material past the ? - the spider sees the ? and stops
 - Spider Traps
 - Sites generated dynamically from a database (.cfm, asp, .cgi)
 - Different pages crawled by different engines (no overlap) - [no two engines are alike](#)
 - Real-time content - spiders do not crawl/recrawl in real-time (too much information, no real good reason to spider this type of information (stock quotes, weather, airline flight arrival/departure information)).
-

Why Use Invisible Web Sources?

- Specialized content focus - more comprehensive results - subject specific
 - Specialized Search Interface - more control over search input and output
 - Increased Precision and Recall
 - Invisible Web Resources - highest level of authority
 - Answer may not be available anywhere else.
-

When to Use an Invisible Web Resource

- When you are familiar with a subject.
 - when you are familiar with a specific search tool
 - When you are looking for a precise answer.
 - when you want authoritative, exhaustive, results.
 - When timeliness of content is an issue.
-

Recent article in Chronicle of Higher Education "[New Allies in the Fight Against Research By Googling: Faculty Members and Librarians Slowly Start to Work together on Courseware.](#)"
(March 21, 2003)

- Google is not the only research tool available
 - Subject Specific Databases
 - Highly authoritative research tools
 - Information is already organized for optimal Results
-

Invisible Web Resources

Examples of Invisible Web Tools

[Well Connected](#)

The Center for Public Integrity has developed a searchable database containing ownership information for radio, tv, cable stations, and telephone companies in US. The database is searchable by company name, geographic area, and call sign.

[The Kaiser Commission on Medicaid and the Uninsured](#)

Medicaid Benefits: Services Covered, Limits, Copayments and Reimbursement Methodologies for 50 states, DC, and the Territories (as of January 2003)

[Internet Movie Database](#) | [All Music Guide](#)

[US Patent Databases](#)

Numerous search options including full text and bibliographic databases. Full text of all US patents issued since January 1, 1976, and full-page images of each page of every US patent issued since 1790.

[Mind - The Meetings Index](#)

The Meetings Index offers free access to locate future conferences, congresses, meetings, and symposia. In the areas of science/technology, medical/life sciences, pollution control/ecology, and social sciences/humanities

[ClinicalTrials.gov](#)

The US National Institutes of Health, through its National Library of Medicine, has developed current information about clinical research studies.

[SMEALSearch](#)

"Search engine that searches web and catalogs for academic articles that are in any area of business.

Specialized Search tools

Video/Audio Search

[PBS NewsHour with Jim Lehrer](#) - Keyword Search, Watch Video/List Audio

National Public Radio – [Audio Archives](#)

[Speechbot](#) - search for radio programs by keyword, searches across radio programs (uses speech recognition software to create a transcript of the program and then builds an index of the words spoken during the program).

[CapitolHearings.org](#) - Listen/Watch Senate Hearings – (Live)

Web Archives

[The WayBack Machine](#)

The WayBack Machine, a service from the Internet Archive and Alexa Internet, allows people to access and use archived versions of stored websites. Visitors to the Wayback Machine can type in an URL, select a date, and then begin surfing on an archived version of the web. The WayBack Machine is built so that it can be used and referenced by anybody and everybody. [Recall Search](#) is now available.

[Amazon - Search Inside the Book](#)

Specialized Web Search Resources by Major Web Search Engines

[AlltheWeb](#)

News
Multimedia Catalogs
Pictures
Videos
Audio Files
FTP Files

[Altavista](#)

Images
MP3/Audio
Video
Directory
News

[Google](#)

Images
Groups
Directory
News

[Froogle](#)

[Google Catalogs](#)
[Unclesam](#)

[Teoma](#)

(Ask.com, Lycos, Metacrawler, Infospace, and Excite)

[Yahoo](#)

Images
Directory
Yellow Pages
News
Products
Maps

Keeping Current: Resources to Monitor

[ResourceShelf](#)

Gary Price - daily updates on new useful tools, resources, databases, and search engine news.

[Librarians' Index to the Internet](#)

"The Librarians' Index to the Internet is a searchable, annotated subject directory of more than 12,000 Internet resources selected and evaluated by librarians for their usefulness to users of public libraries."

[InfoMine](#)

InfoMine is a virtual library of Internet resources

[Marylaine Block's Neat New Stuff I found on the Web this Week](#)

[Free Pint](#)

Email newsletters - Will Hahn - (Monthly)

Internet Resources Newsletter (Monthly)

Free, monthly newsletter for academics, students, engineers, scientists, and social scientists

[The Virtual Chase](#)

Genie Tyburski's Daily Updates on web resources

[LLXR](#) ||| [LLRX Buzz](#) ||| [BeSpecific](#)

Law, Technology, Internet Resources (Sabrina Pacific)

[Scout Report & Scout Report Archive](#)

University Of Wisconsin, Madison - weekly current awareness newsletter

[Searchenginewatch.com](#)

Danny Sullivan and Chris Sherman - useful articles on search engines

[Search Engine Showdown](#)

Greg R. Notess - Evaluates search engine features

[NoodleBib - Best Search Engine for your Needs](#)

Debbie Abilock's evaluation of search engines

The Need for Specialized Tools & Knowledge of the Invisible Web

- The web and general web databases will continue to grow larger
- Existing and new specialized databases will be released and made available
- To improve the chances of finding information - these specialized databases will increase in importance
- In many cases specialized tools, invisible and specialized tools will have interfaces “customized” for the [specific data in the database](#). Example - Government Accounting Office advanced search
- Ability to sort, view data in ways specific to the data set - use the search tool that is designed to search the database or material - you will have much greater control over the search and hopefully you will have better results.
- Bigger databases translates - more recall, lower precision (more pages, not necessarily results that are on target)
- Focuses databases, smaller universe of materials to search
- Greater ability to “work with the data” (field searching, sorting, limits)
- The authority of the author will increase in importance, you know where the information is coming from
- Use the right tool for the job - encyclopedia, phone numbers, etc...
- Think Resources “learn” them like you learn traditional reference tools
- LexisNexis, Dialog - offer many databases depending on the information you are looking for - and resources available to you
- Deciding where and what to search is a skill that information professionals have

Conclusions

- Be Aware of the limitations of general search engines
- No single engine indexes the entire web
- Use more than one search engine
- Even if it “might” be accessible in a general engine would a focused engine get it to you more quickly?
- The challenge is learning the many different resources available (both visible and invisible and being able to access it quickly)
- Web collection development will be more important
Building a collection, knowing what’s available - both visible and invisible tools
- Think of specialized and invisible web tools like you think of your reference collection
- Future? [federated searching](#), [broadcast searching](#) - searching multiple databases at one time
([Science.gov](#))
- For our end users, products like MuseGlobal offer great promise
- Can handle any database, merge results into a single list, remove duplicates, customized for each library

